NOVÝ ARCHIV BETON TKS JE KOMPLETNÍ!

Do archivu časopisu Beton TKS na webových stránkách ***www.betontks.cz*** byl v polovině května tohoto roku doplněn poslední chybějící výtisk – historicky první výtisk z roku 2001. Archiv je nyní kompletní a obsahuje více než 2200 článků uveřejněných během celé doby vydávání časopisu. Jak archiv vznikal?

Časopis je vydáván již sedmnáctým rokem, a tak objem publikovaných článků nabyl již úctyhodných rozměrů. Původní webové stránky obsahovaly jak archiv, tak vyhledávání, ale pouze v omezené podobě. Když došlo k jejich modernizaci, bylo nejdůležitější částí zadání vytvořit archiv, ve kterém lze a rychle vyhledávat, tak aby odpovídal 21. století. Zadání jasné, stručné. Samotné projektování a programování databáze archivu a vytvoření webového rozhraní se ukázalo být nejjednodušší částí projektu. Problém přišel s naplněním databáze vhodnými daty.

Většinu čísel časopisu jsme měli ve formě pdf souborů, kromě prvního ročníku, kdy se tiskové podklady ještě připravovaly „analogovou“ cestou. Každé číslo se muselo digitálně rozstříhat na příslušné články. Po této fázi jsme měli v ruce přes 2000 článků. A zde se při umístění na web objevil problém.

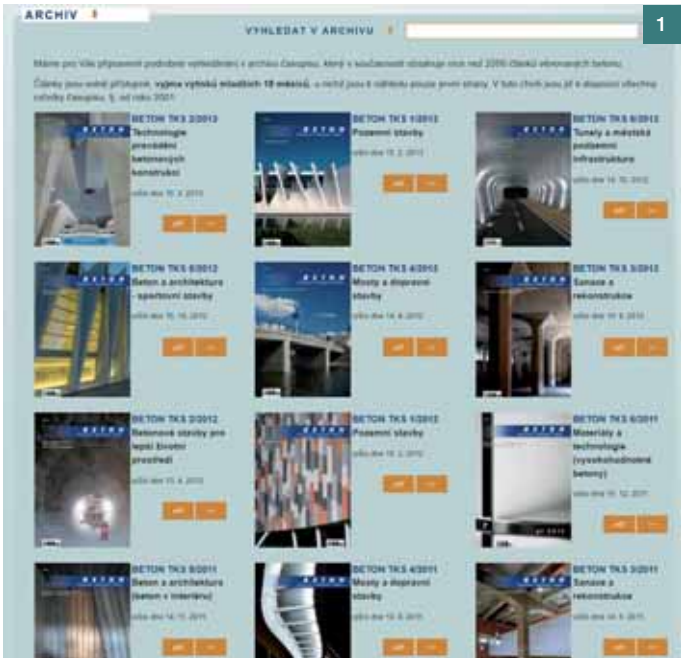
Ve světě informačních technologií je 17 let dlouhá doba. Za ten čas prošel časopis několikrát změnou layoutu a samotný formát pdf se také velmi výrazně proměnil. Výsledkem bylo, že webové prohlížeče neuměly správně rozlišovat a vyhledávat texty v pdf souborech. Tím bylo automatické vyhledávání zcela vyloučeno. Nějakou dobu jsme zkoušeli vyhledávání pomocí algoritmů společnosti Google, ale ani tato společnost si uspokojivě neporadila se starými formáty našich souborů. Chybovost při vyhledávání textů byla příliš veliká, tím víc, čím hlouběji do minulosti jsme se dostávali. Nesnáze byly nejvíce znát mezi lety 2004 a 2005. V pdf souborech z roku 2004 a starších úplně selhávalo rozpoznávání slov i znaků.

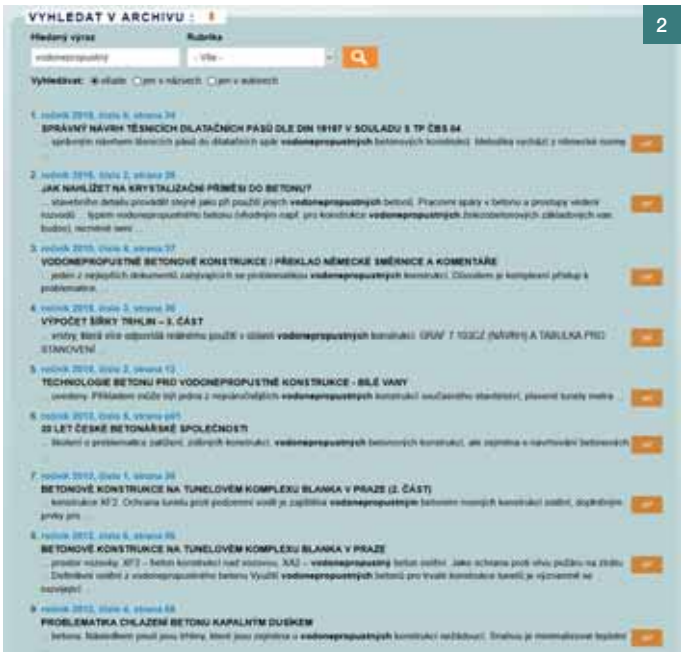
Vyzkoušeli jsme i metodu OCR, kdy se jednotlivé soubory netváří jako dokumenty, ale jako běžný obrázek. V tomto obrazu se snaží „inteligentní“ software rozpoznávat texty. I tak bylo nesprávných převodů znaků mnoho, zejména u některých starších čísel, v nichž byl použit poněkud „roztahaný“ font s velkými mezerami mezi znaky.

Jelikož jsme chtěli mít vyhledávání s co nejmenším počtem chyb, nezbývalo než sáhnout k hrubé síle. Bylo zřejmé, že samotné pdf soubory bohužel nestačí. Ze všech článků bylo třeba „vypreparovat“ textovou část. To se dělalo několika způsoby. U článků do roku 2009 kopírováním pomocí programu Adobe Acrobat, u článků mezi lety 2009 až 2005 pomocí nástroje Google Docs a u starších článků pomocí OCR technologie společnosti ABBYY. I tak musel být poté text vždy ještě ručně zkontrolován – lidská síla je zatím stále nenahraditelná. Až takto připraveným „destilovaným“ textem jsme mohli databázi doplnit.

Dnes je archiv kompletní a připravený pro vyhledávání ve všech článcích, které byly za dobu vydávání časopisu publikovány. Je umožněno vyhledávání v celém obsahu nebo pouze v názvech článků a lze též vyhledávat podle jména autora. K náhledu jsou připravena jak celá čísla, tak jednotlivé články (výtisky starší než 18 měsíců jsou k dispozici v plné verzi, mladší výtisky pouze v náhledu (první strany všech článků)).

Doufáme, že Vám archiv dobře poslouží.





Obr. 1 Úvodní stránka archivu Beton TKS – seřazená vydání a jejich titulní strany, dostupná jsou jak celá čísla, tak jednotlivé články, výtisky starší než 18 měsíců jsou k dispozici v plné verzi, mladší výtisky pouze v náhledu (první strany všech článků)

Obr. 2 Výsledky vyhledávání v archivu na výraz „vodonepropustný“ seřazené podle data vydání od nejmladších článků k nejstarším

P.S. Během projektu jsem učinil dvě zajímavá poznání. Zaprvé, jak i informace uložené v elektronické formě zastarávají. Formát pdf se za ta léta změnil tak, že nová verze programu má problémy se zpracováním dokumentů vytvořených před 15 lety. Zadruhé, že ani společnost Google není všemocná a její nástroje si neporadí se vším (alespoň ne dnes).

Milan Senko
Svaz výrobců betonu ČR
e-mail: milan.senko@svb.cz